



万成瑞,于茹月,盖建勋,徐圣贤,黄元仿.土壤全氮驱动因素分析及模型构建——以湟水河流域为例[J].中国农业大学学报,2025,30(10):25-37.
WAN Chengrui,YU Ruyue,GAI Jianxun,XU Shengxian,HUANG Yuanfang. Soil total nitrogen driving factor analysis and model construction: A case study of the Huangshui River Basin[J]. *Journal of China Agricultural University*, 2025, 30(10): 25-37.
DOI: 10.11841/j.issn.1007-4333.2025.10.03

土壤全氮驱动因素分析及模型构建 ——以湟水河流域为例

万成瑞 于茹月 盖建勋 徐圣贤 黄元仿*

(中国农业大学 土地科学与技术学院/农业农村部华北耕地保护重点实验室,北京 100193)

摘要 为探究土壤氮素的驱动因素,维护湟水河流域生态脆弱区的生态系统稳定,基于湟水河流域土壤全氮(Total Nitrogen, TN)实测数据,采用随机森林、皮尔逊相关性分析和多尺度地理加权回归方法,构建混合模型对土壤TN驱动因素进行分析。结果表明:1)湟水河流域的土壤TN呈现由东南向西北递增的趋势,高值区主要在流域西北的高海拔地区,低值区零星分布在湟水河中下游段的两侧。2)土壤TN的主要驱动因素为年均气温、pH、距城市密集区距离,各因素对土壤TN的影响大小和作用尺度存在差异。特征重要性由高到低为年均气温(78.5%)>pH(6.9%)>距城市密集区距离(1.4%);带宽由高到低为年均气温(278 km)>距城市密集区距离(85 km)>pH(48 km)。3)原始模型采用变量个数为19,混合模型在仅采用3个变量的基础上,达到了相等的校正 R^2 (0.41)。同时,混合模型在稳定性指标AICc和有效自由度上均优于原始模型AICc和有效自由度。综上,湟水河流域土壤TN总体呈现出带状的空间分布格局,年均气温是影响土壤TN的最主要自然因素,其影响主要体现在较大尺度上;人类活动对土壤TN的影响较小,且影响主要集中在城市周边区域;混合模型有效减轻了过拟合现象,在该流域适用性良好。本研究可为湟水河流域的农业可持续发展及生态保护提供科学依据。

关键词 湟水河流域;土壤全氮;驱动因素;混合模型;随机森林;多尺度地理加权回归

中图分类号 S159.2 **文章编号** 1007-4333(2025)10-0025-13 **文献标志码** A

Soil total nitrogen driving factor analysis and model construction: A case study of the Huangshui River Basin

WAN Chengrui, YU Ruyue, GAI Jianxun, XU Shengxian, HUANG Yuanfang*

(College of Land Science and Technology/Key Laboratory of Arable Land Conservation in North China of Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing 100193, China)

Abstract To better understand the drivers of soil nitrogen, maintain the stability of the ecological system in the Huangshui River Basin, this study based on a set of measured soil Total Nitrogen (TN) data in the Huangshui River Basin, a hybrid model was developed using the random forest, Pearson correlation analysis, and multi-scale geographically weighted regression methods was adopted to analyze the driving factors of soil TN. The results show that: 1) Soil TN exhibited a gradual increase from the southeastern to northwestern regions of the basin. High-value areas were concentrated in the northwestern high-altitude zones, while low-value areas sporadically occurred along the mid-to-lower reaches of the Huangshui River. 2) The primary drivers were the annual mean temperature, pH, and distance to urban areas. There were differences in the magnitude and scale

收稿日期: 2025-05-07

基金项目: 国家自然科学基金项目(U20A20115)

第一作者: 万成瑞(ORCID:0009-0009-7072-2918),博士研究生,E-mail:bs20243211300@cau.edu.cn

通讯作者: 黄元仿(ORCID:0000-0002-6706-0361),教授,博士生导师,主要从事数字农业与土壤过程量化研究,E-mail:yfhuang@cau.edu.cn

of the impact of various factors on soil TN. Feature importance from highest to lowest is: Annual mean temperature (78.5%)>pH (6.9%)>distance to urban areas (1.4%). Bandwidth from highest to lowest is: Annual mean temperature (278 km)>distance to urban areas (85 km)>pH (48 km). 3) The hybrid model, using only three variables, achieved equivalent adjusted R^2 (0.41) to the original 19-variable model. It also outperformed the original model in terms of stability index AICc and effective degrees. The soil TN in the Huangshui River Basin displayed a belt-shaped spatial distribution pattern. Annual mean temperature, operating at larger spatial scales, was the dominant natural driver, while human activities exerted minor localized effects near urban areas. The hybrid model effectively reduced overfitting and has good applicability in this basin. This study provides a scientific reference for sustainable agriculture and ecological protection in the Huangshui River Basin.

Keywords Huangshui River Basin; soil total nitrogen; driving factors; hybrid model; random forest; multiscale geographically weighted regression

土壤氮素是植物生长必需的营养元素之一,是衡量土壤质量和肥力的关键指标。然而,土壤中过量的氮素则会引发土壤质量下降、水体富营养化、大气污染等环境问题,对生态系统的稳定性造成威胁^[1-2]。因此,了解和掌握土壤氮素含量的空间分布及其驱动因素,对合理调控土壤氮素平衡、实现农业可持续发展和生态保护具有重要意义^[3]。20世纪末期,关于驱动因素的研究主要依赖于最小二乘法回归(Ordinary Least Squares, OLS)及衍生算法,假设自变量和因变量之间具有线性关系,误差项相互独立并服从同一分布,且误差项具有同方差性和正态分布^[4-5]。而在实际应用中,土壤养分与其驱动因素之间的关系往往更加复杂,存在非线性、多重共线性等问题,这些问题使得OLS的应用受到了限制^[6-8]。随着计算能力的提升和大数据技术的应用,机器学习方法被广泛应用于环境科学领域的研究中^[9-10]。随机森林(Random Forest, RF)作为一种集成学习方法,由Breiman^[11]在2001年首次提出。随机森林通过随机选择特征和数据子集来生成每棵决策树,并利用投票机制来得到最终预测结果,具有准确性高,抗噪声能力强,能够处理非线性和高维数据等优点^[9]。近年来,随机森林因其出色的预测能力和抗噪声能力,逐渐被广泛应用于土壤科学领域的空间预测和驱动因素分析中,例如:卢宏亮等^[12]对安徽省土壤养分属性进行了预测,结果表明随机森林模型具有较高的预测精度;高文静等^[13]利用随机森林成功识别了影响土壤有机碳的重要因素。除此之外,在其他土壤属性、特征优化等方面的研究均利用随机森林取得了较好的效果^[14-18]。

然而,随机森林虽然能很好地拟合多组数据,但该方法并不基于变量之间的空间关系进行解释^[19],Verikas等^[20]也指出,随机森林结果并非在任

何情况下都可采信。这限制了其在地理空间数据分析中的应用。为了弥补随机森林对于存在空间关系变量的解释性不足,可以将多尺度地理加权回归(Multiscale Geographically Weighted Regression, MGWR)方法引入到土壤养分的驱动因素分析中。MGWR是对地理加权回归(Geographically Weighted Regression, GWR)的扩展^[21],由Fotheringham等^[22]在2017年提出。MGWR进一步考虑了不同尺度下的空间异质性,使得模型参数能够随空间位置的变化而变化,从而更好地捕捉变量之间的空间关系,能够对具有空间异质性的变量进行更加精确的拟合^[23]。例如:沈体雁等^[24]展示了MGWR在解释复杂空间现象中的优势;南富森等^[25]利用MGWR得出了影响黄河流域土壤有机碳的主要因素;赵明松等^[26]研究表明,在MGWR、Mixed GWR、GWR和MLR 4种模型中,MGWR具有最优精度。但在实际运用中,MGWR在处理大数据时面临着计算成本过高和数据过拟合的问题,在使用较多环境变量进行分析时,由于变量间关系复杂,模型的精度往往不高。

湟水河流域位于青藏高原,生态系统敏感、脆弱,环境自净能力有限^[27]。因此,了解和掌握该流域土壤氮素含量的空间分布及其驱动因素,对合理调控土壤氮素平衡、实现农业可持续发展和生态保护具有重要意义^[28]。因此,探究土壤氮素的驱动因素,维护湟水河流域生态脆弱区的生态系统稳定,本研究拟以湟水河流域为研究对象,采用随机森林、皮尔逊相关性分析和多尺度地理加权回归方法,对土壤TN的空间分布特征及其驱动因素进行综合分析。通过构建筛选模型识别影响土壤TN的关键驱动因素,分析各驱动因素在不同空间尺度上的作用强度和范围,对比混合模型与传统方法的性

能差异,本研究以期湟水河流域土壤氮素的科学管理、农业可持续发展以及生态环境保护提供理论支持和技术指导,为空间分析研究提供新的思路与方法。

1 材料与方法

1.1 研究区概况

湟水河流域位于黄河上游,地处 36°02′ N~37°28′ N, 100°41′ E~103°04′ E, 跨海北藏族自治州(海晏县)、西宁市(大通县、湟源县、湟中区、城北区、城西区、城中区、城东区)、海东市(互助县、平安

区、乐都区、民和县)3市12个区县,总面积约为 25 400 km²[29]。湟水河流域南邻青藏高原,北夹黄土高原和达坂山,形成西北高而东南低的三角地带,地形地貌复杂多样,海拔在 1 605~4 881 m,土壤类型主要以栗钙土、黑毡土、石灰性黑钙土和黑钙土为主;属高原山地气候,全年低温,降水较少,年均气温在 2.7~7.8℃,年平均降水量 486.6 mm。湟水河流域的面积占青海省土地总面积的 2.3%,但人口约占全省的 56%,耕地面积约占 50%,工农业生产总值约占 54.2%^[30],是青海省的政治、经济、文化中心。



图1 研究区域及样点分布

Fig. 1 Study area and sampling points

1.2 数据来源

湟水河流域的高程数据来源于美国地质调查局),分辨率为 30 m(<https://lpdaac.usgs.gov/products/astgtmv003/>);土壤类型图来源于国家地球系统科学数据中心土壤分中心(<http://soil.geodata.cn/>);土地利用现状图、降水、气温数据均来源于中国科学院资源环境科学数据平台(<http://www.resdc.cn/>),研究区内气温依据气象站点的多年平均气温数据空

间插值获得,距城市密集区距离依据城市建成区边界计算获得;河网数据来源于国家地球系统科学数据共享平台(<http://www.geodata.cn/>);土壤全氮含量、农田林网化程度、耕层厚度、清洁程度等数据来源于 2020 年农业农村部测土配方施肥数据,该数据包括 591 个表层土壤(0—20 cm)样品,采用 pH 计测定 pH, $m(\text{水}):m(\text{土})$ 为 2.5:1。采用凯氏定氮法测定土壤全氮含量^[31]。

1.3 研究方法

1.3.1 经验贝叶斯克里金法

经验贝叶斯克里金法 (Empirical Bayesian Kriging) 是一种空间插值方法, 由 Krivoruchko^[32] 于 2012 年提出。该方法在传统克里金法的基础上, 结合了贝叶斯统计的思想, 通过多次迭代结果来估计半变异函数, 以更好地解释半变异函数参数的不确定性。经验贝叶斯克里金法在 ArcGIS Pro 3.0.0 的统计向导模块中提供, 可自动执行构建克里金模型所需步骤。

1.3.2 全局莫兰指数

全局莫兰指数 I_{global} (Global Moran's I) 是衡量空间数据全局自相关性的重要统计指标。其计算公式如下:

$$I_{\text{global}} = \frac{n \sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\left(\sum_i \sum_j w_{ij} \right) \sum_i (x_i - \bar{x})^2} \quad (1)$$

式中: n 为采样点数量; w_{ij} 为采样点 i 与 j 间的空间权重矩阵; x_i 和 x_j 分别为采样点 i 与 j 处的观测值; $\bar{x}$ 为观测值的平均值。该指数可以检测研究区域内空间变量是否存在显著的空间聚集性。全局莫兰指数的 $P < 0.01$ 代表结果在 99% 的置信区间显著, $P < 0.05$ 代表结果在 95% 的置信区间内显著。若不满足显著性检验, 则变量在研究区域的空间自相关性需要谨慎解释。全局莫兰指数的值介于 $[-1, 1]$, 当指数接近 1 时, 表明变量具有显著的正相关性, 即高值或低值倾向于在空间上聚集; 当指数接近 -1 时, 表明变量具有显著的负相关性, 即高值和低值在空间上交替出现; 而当指数接近 0 时, 表明变量在空间上随机分布, 没有显著的空间自相关性。

1.3.3 局部莫兰指数

局部莫兰指数 I_{local} (Local Moran's I) 可以识别变量在局部区域的空间聚集模式。其计算公式如下:

$$I_{\text{local}} = \frac{(x_i - \bar{x})}{S^2_x} \sum_j [w_{ij} (x_j - \bar{x})] \quad (2)$$

式中: w_{ij} 为采样点 i 与 j 间的空间权重矩阵; x_i 和 x_j 分别为采样点 i 与 j 处的观测值; $\bar{x}$ 为观测值的平均值。局部莫兰指数能够通过计算每个样本点的局部自相关性, 识别出空间数据中的高值点、低值点以及异常点, 即高高聚集 (HH) 高值区域周围邻接高值区域; 低低聚集 (LL): 低值区域周围邻接低值区域; 高低聚集 (HL): 高值区域周围邻接低值区域;

低高聚集 (LH): 低值区域周围邻接高值区域。

1.3.4 皮尔逊相关性分析

皮尔逊相关性分析 (Pearson correlation analysis) 是一种统计方法, 用于评估 2 个连续变量之间的线性相关程度。其计算公式如下:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (3)$$

式中: r 为皮尔逊相关系数; x_i 和 y_i 分别为变量 x 和 y 的观测值; $\bar{x}$ 和 $\bar{y}$ 分别为变量 x 和 y 的平均值。皮尔逊相关性分析能够通过计算 2 个变量之间的相关系数, 识别出变量间的线性关系强度和方向。具体来说, 当相关系数接近 1 时, 表示 2 个变量之间存在强正相关关系, 即一个变量的增加往往伴随着另一个变量的增加; 当相关系数接近 -1 时, 表示 2 个变量之间存在强负相关关系, 即一个变量的增加往往伴随着另一个变量的减少; 当相关系数接近 0 时, 表示 2 个变量之间没有显著的线性相关性。

1.3.5 随机森林

随机森林是一种通过集成多棵决策树进行分类和回归的机器学习算法。随机森林的预测值通过集成所有决策树的预测结果得到, 具体计算公式如下:

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^N T_i(x) \quad (4)$$

式中: $\hat{f}(x)$ 为样本 x 的最终预测值; N 为随机森林中决策树的数量; $T_i(x)$ 为第 i 棵决策树对样本 x 的预测结果。随机森林通过引入随机性和集成学习的思想来提高模型的稳定性和预测性能, 可以有效处理高维数据, 并能评估各个特征的重要性。本研究该部分的计算基于 SPSSPRO 1.1.25 软件平台, 输入变量如表 1 所示。所有定类变量均经过哑变量方式转为虚拟变量输入。

1.3.6 多尺度地理加权回归

多尺度地理加权回归是在传统地理加权回归基础上发展起来的一种更为灵活和精准的空间回归方法。MGWR 通过为每个解释变量引入不同的带宽参数, 能够在多个尺度上捕捉变量之间的空间关系, 适于研究具有空间异质性的变量。MGWR 的基本公式如下:

$$y_i = \beta_{\text{bw0}}(u_i, v_i) + \sum \beta_{\text{bwk}}(u_i, v_i) x_{ik} + \epsilon_i \quad (5)$$

式中: y_i 为因变量在观测点 i 的值; $\beta_{\text{bw0}}(u_i, v_i)$ 为带宽 bw0 时的回归截距; $\beta_{\text{bwk}}(u_i, v_i)$ 为最优带宽 bwk 下的第

表 1 本研究变量描述

Table 1 Descriptions of explanatory in this study

变量名称 Variable name	单位 Unit	变量描述 Variable description
年均气温 Annual mean temperature	℃	年平均气温
pH		土壤酸碱度
纬度 Latitude	(°)	样点位置的纬度
年均降雨量 Annual mean precipitation	mm	年平均降雨量
海拔 Elevation	m	海拔
距城市密集区距离 Distance to urban concentrated area	km	距流域内最大城市密集区距离
经度 Longitude	(°)	样点位置的经度
有效土层厚度 Effective soil layer thickness	cm	有效土层厚度
耕层土壤容重 Bulk density of plough layer soil	g/cm ³	耕层土壤的密度
耕层厚度 Plough layer thickness	cm	耕层土壤厚度
土壤类型 Soil type	虚拟变量	分类为黑钙土、栗钙土、潮土、新积土、灰钙土、灌淤土
母质类型 Soil parent material type	虚拟变量	分类为红土母质、黄土母质、混合物
作物类型 Crop type	虚拟变量	分类为粮食作物、非粮食作物
灌溉能力 Irrigation capacity	虚拟变量	分类为不满足、基本满足、充分满足
生物多样性 Soil biodiversity	虚拟变量	分类为丰富、不丰富
农田林网化率 Farmland shelter rate	虚拟变量	分类为高、低
耕层质地 Plough layer soil texture	虚拟变量	分类为砂土、壤土、黏土
清洁程度 Soil cleanliness degree	虚拟变量	分类为清洁、尚清洁
熟制 Cropping system	虚拟变量	分类为一年一熟、一年两熟

k 个变量的回归系数; x_{ik} 为变量 x_k 在观测点 i 的值; ϵ_i 为随机误差项。该部分的计算基于 ArcGIS Pro 3.0.0 软件平台。

2 结果与分析

2.1 湟水河流域土壤全氮空间分布特征

研究区土壤全氮描述性统计及空间插值精度验证见表 2。可见:研究区土壤 TN 的全局莫兰指数为 0.20,标准化 Z 值为 11.30,大于显著性检验的临界值

2.56($P<0.01$),表明土壤 TN 在研究区内存在显著的空间自相关性;土壤 TN 的变异系数分别为 0.35,表明土壤 TN 具有中等强度的变异性;研究区土壤 TN 的分布类型呈对数正态分布,偏度为 1.04,丰度为 2.29。基于全局空间自相关性分析,对湟水河流域的土壤 TN 采用经验贝叶斯克里金法进行空间插值,利用局部莫兰指数法识别高值和低值的聚集区;土壤 TN 空间插值的 SRMSE 值接近于 1,表明模型的预测误差与实际数据的变异程度相当。

表2 湟水河流域土壤TN描述性统计及空间插值精度验证

Table 2 Descriptive statistics and spatial interpolation accuracy validation for soil TN in the Huangshui River Basin

指标 Index	全氮 Total nitrogen	指标 Index	全氮 Total nitrogen
平均值 Mean	1.51 g/kg	Z值 Z value	11.30
标准差 Standard deviation	0.52 g/kg	P值 P value	<0.01
变异系数 Coefficient of variation	0.35	空间插值函数	去趋势
分布类型 Distribution type	对数正态	Spatial interpolation function	K-Bessel
偏度 Skewness	1.04	均方根误差 RMSE	0.47 g/kg
峰度 Kurtosis	2.29	连续排序概率得分 CRPS	0.25 g/kg
全局莫兰指数 Moran's I	0.20	标准化均方根误差 SRMSE	1.02

图2和图3分别为湟水河流域土壤TN的空间分布图和局部莫兰指数图。由图2可见土壤TN的空间分布呈现由东南向西北递增的趋势,高值区主要分布在流域西北的高海拔地区,低值区分布在湟水河中下游段的

两侧。由图3可见局部莫兰指数图进一步揭示了土壤TN的空间聚集模式和局部空间自相关性。土壤TN的空间聚集程度较高,且高值区和低值区之间存在较为明显的过渡带,整体呈现出带状的空间分布格局。

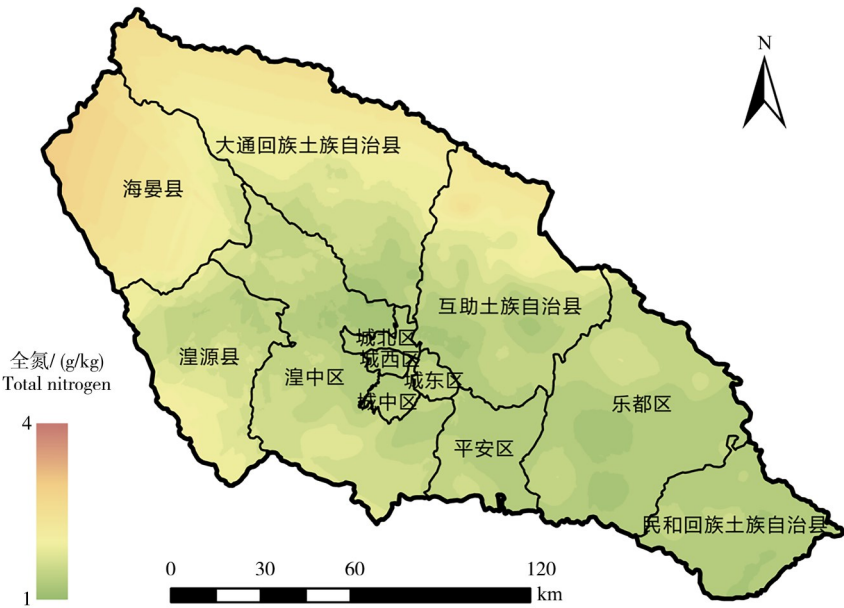


图2 湟水河流域土壤TN空间分布图

Fig. 2 Spatial distribution of soil TN in the Huangshui River Basin

2.2 湟水河流域土壤全氮的主要驱动因素

图4为土壤TN各驱动因素的特征重要性排序结果。可见特征重要性>1%的驱动因素包括:年均气温(78.5%)、pH(6.9%)、纬度(4.0%)、年均降水量(2.8%)、海拔(2.2%)距城市密集区距离(1.4%),表明自然因素中的年均气温是影响湟水河流域土壤TN的最主要因素,人为因素影响较小。

图5为土壤TN驱动因素相关性分析热图,该部分选取特征重要性>1%的因素进行皮尔逊相关性分析。可见:年均气温与纬度呈强负相关(−0.51),与海拔呈极强负相关(−0.92),与年均降水量呈较强负相关(−0.78);pH与多数变量相关性较低(最大绝对值为0.17),可能表明土壤酸碱度受多种复杂因素影响,其他环境变量难以解释其变化;距城市密集区距

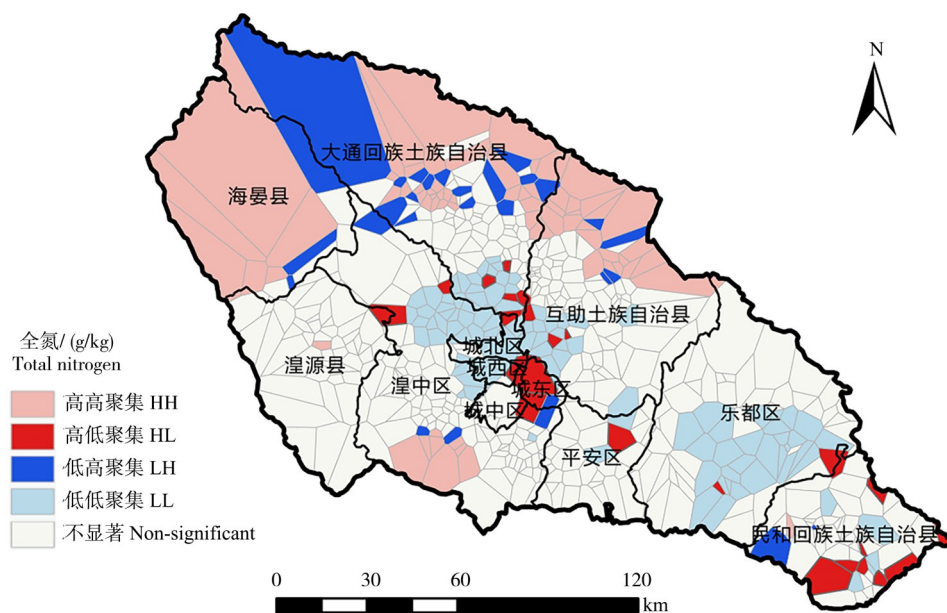


图3 湟水河流域土壤TN局部莫兰指数分布图

Fig. 3 Local Moran's I of soil TN in the Huangshui River Basin

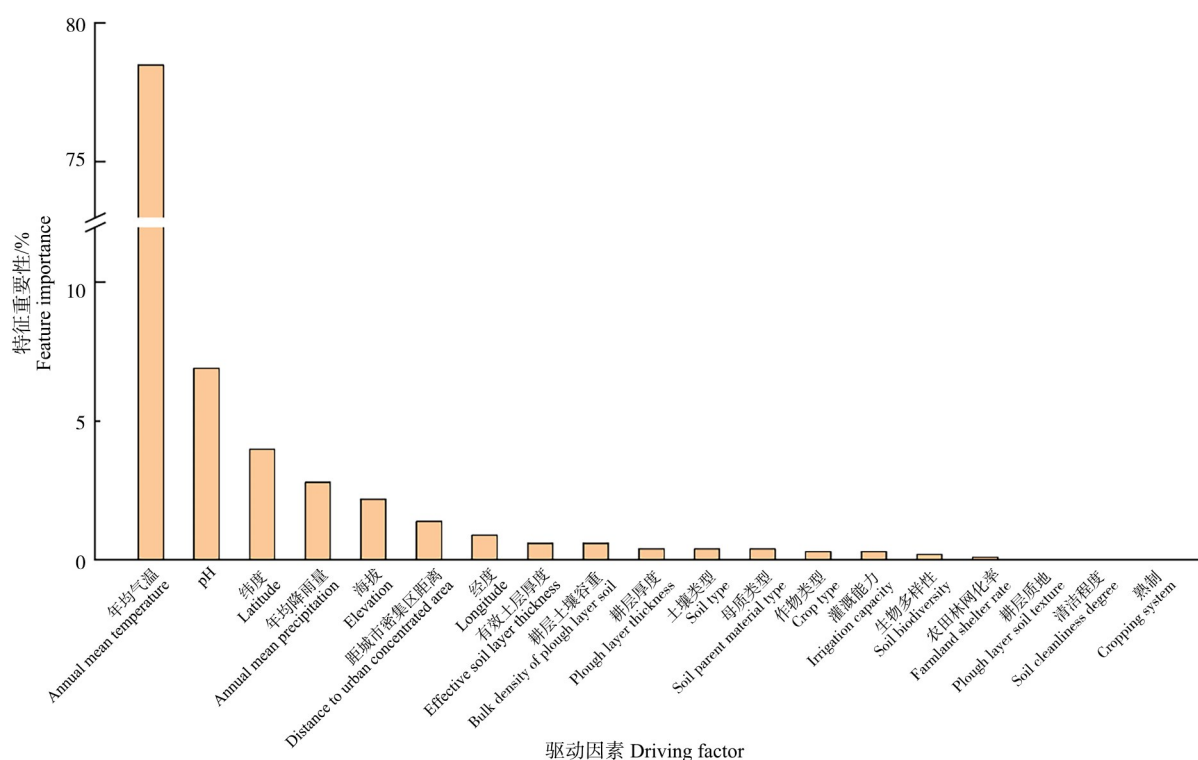


图4 土壤TN驱动因素的特征重要性排序

Fig. 4 Importance ranking of influencing factors for soil TN

离与其他因素的相关性普遍较低(最大绝对值0.36),表明其可能受人类活动或土地利用类型影响,该因素具有一定的独立性。

由于年均气温、纬度、年均降水量与海拔这4个因

素相互之间存在着较强的相关性,且年均气温的特征重要性更高。本研究剔除纬度、年均降水量与海拔3个因素,最终确定以年均气温、pH以及距城市密集区距离作为土壤TN的主要驱动因素,用于后续MGWR

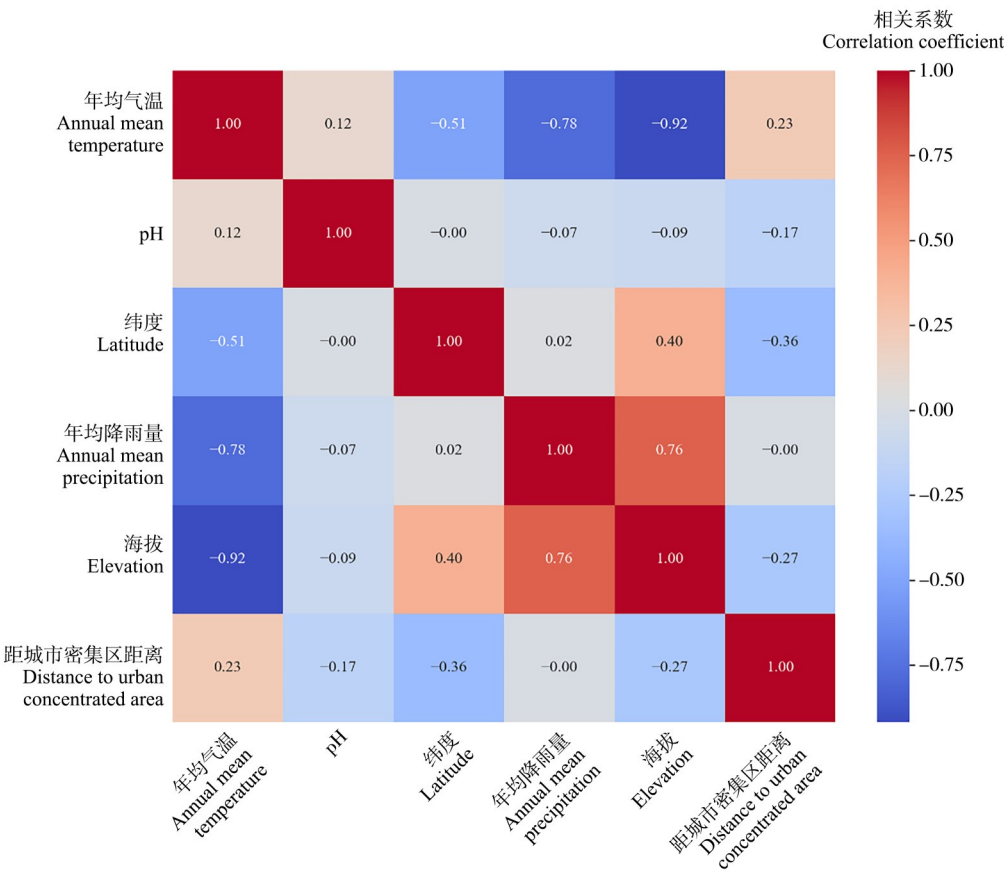


图5 土壤TN驱动因素相关性分析热图

Fig. 5 Heatmap of correlation of soil TN driving factors

建模分析。

混合模型和原始模型的精度、复杂度、鲁棒性参数比较见图6。可见:原始模型的 R^2 和 Sigma^2 分别为0.45和0.55,略高于混合模型0.43和0.57,表明其对数据的拟合精度稍好;原始模型的 R^2 和校正 R^2 差距较大,提示其可能存在过拟合,鲁棒性较差。此外,原始模型使用了全部19个变量,混合模型仅用了3个变量,显著降低了模型复杂度;混合模型的有效自由度为566.23,高于原始模型的551.97,表明混合模型能够更有效地捕捉变量之间的内在关系。AICc是用于衡量模型拟合精度与复杂度平衡的指标,值越小表示模型越好。混合模型的AICc为1375.86,低于原始模型的1389.14,表明混合模型在平衡拟合精度和复杂度方面更优。综合来说,混合模型在复杂度和拟合精度的平衡上表现更好,避免了过拟合,具有更好的泛化能力。

不同驱动因素的空间格局与带宽结果见图7。从各驱动因素的回归系数来看,年均气温的回归系

数均为负值(−0.28~−0.30),表明年均气温与土壤TN呈负相关,即气温越高,土壤TN值越低,且年均气温回归系数的绝对值在研究区南部和中部部分区域较大,表明这些区域气温对土壤TN的影响更强;pH的回归系数均为负值(−0.15~−0.77),变化较大的区域主要集中在研究区的中部和南部;距城市密集区距离的回归系数大部分为正值(0.42~−0.06),表明距城市密集区越远,土壤TN值越高;截距是综合了所有未被显式包含在模型中的因素对因变量的影响,截距的回归系数(0.31~−0.31)代表在所有驱动因素取均值时土壤TN的背景值。

从各驱动因素的带宽来看,年均气温的带宽相对较大(278 km),表明年均气温的影响在空间上具有一定的稳定性,其对土壤TN的作用机制体现在较大的尺度上;pH的带宽相对较小(48 km),表明pH对土壤TN的影响具有较强的空间异质性,不同区域可能因母质、成土过程或人为酸化/碱化措施

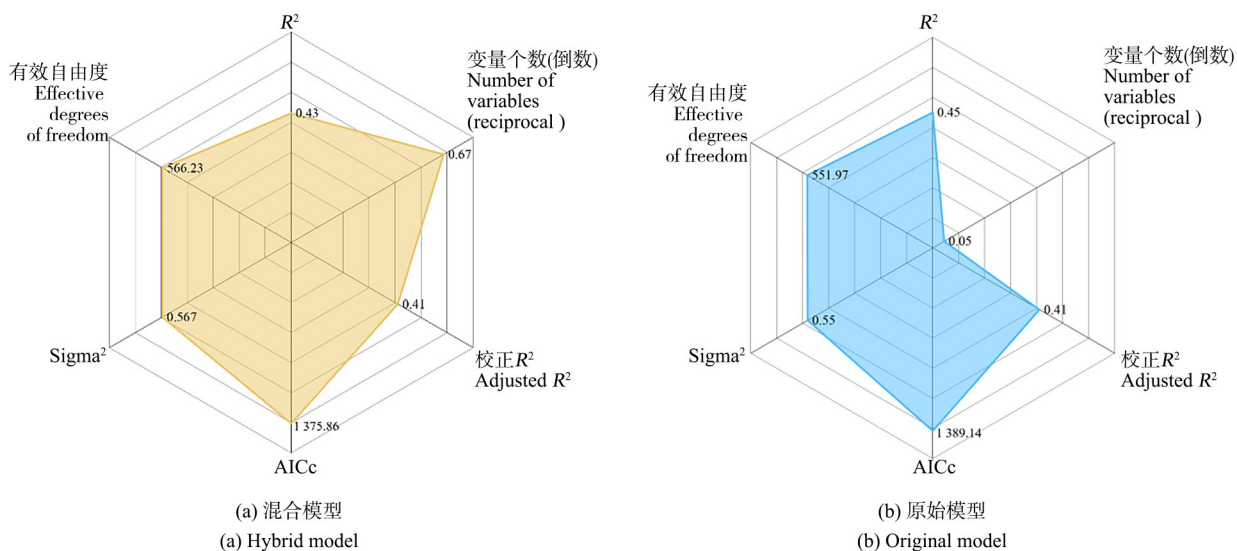


图 6 混合模型与原始模型参数对比

Fig. 6 Comparing performance of mixed and original models

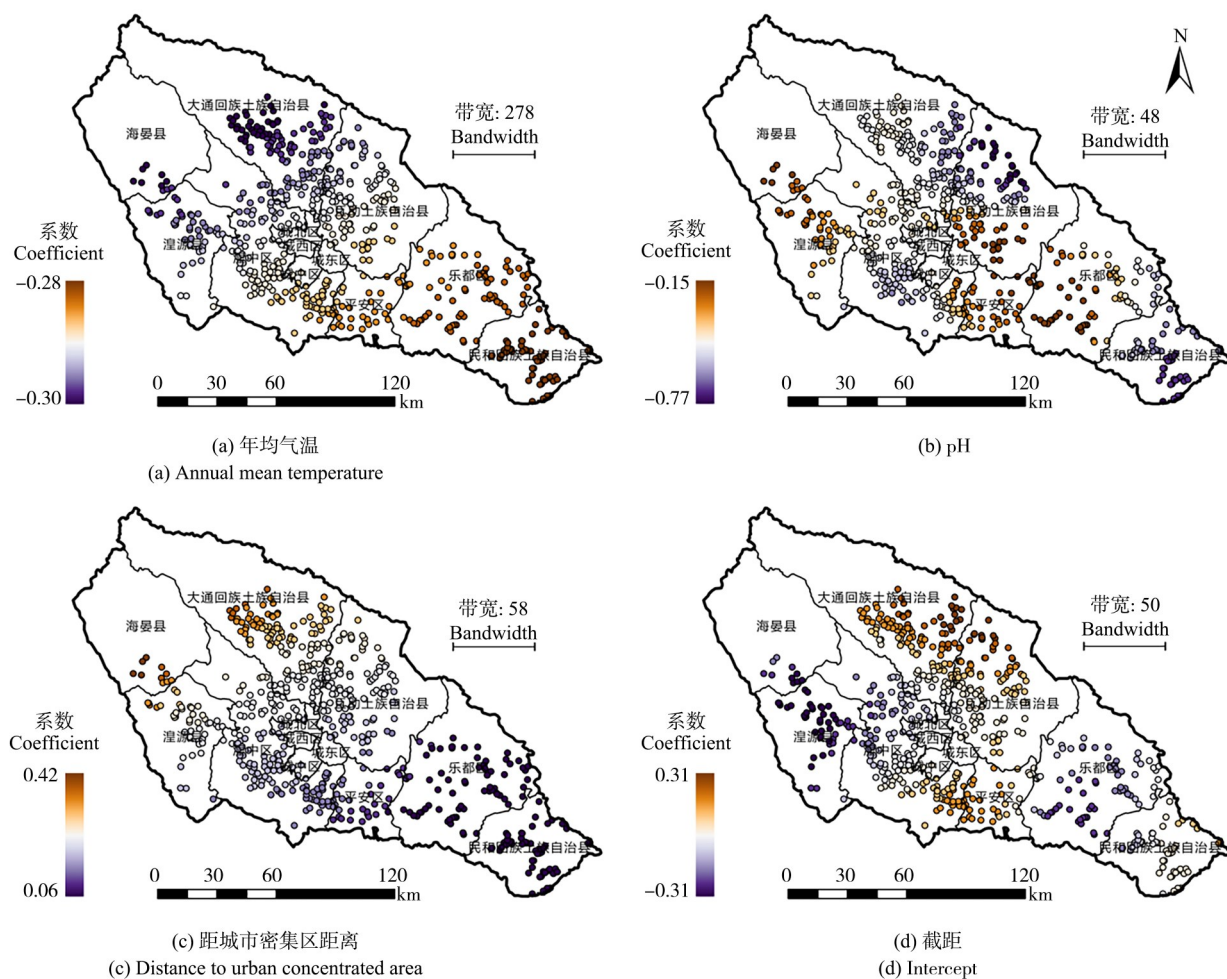


图 7 多尺度地理加权回归系数的空间格局与带宽

Fig. 7 Spatial patterns of coefficients and bandwidth in the MGWR

差异而呈现不同的作用强度;距城市密集区的带宽为85 km,表明城市密集区对土壤TN的影响在一定尺度内具有稳定性,但不同小区域仍存在差异。

总之,年均气温、pH和距城市密集区距离共同塑造了土壤TN的空间格局,而带宽的大小则揭示了各驱动因素在空间尺度上的差异。年均气温的影响较为稳定,适合在较大的范围上进行分析,pH和距城市密集区距离则需要在更精细的空间尺度上进行研究,以揭示其复杂的作用机制。

3 讨论

3.1 湟水河流域土壤全氮的主要驱动因素

湟水河流域的土壤TN分布主要呈西北高而东南低的分布格局,从整体上看具有条带状的分布特征。基于随机森林结果,气温具有最高的特征重要性(78.5%),说明了气温对土壤TN含量的重要影响,MGWR建模结果则进一步说明,气温对土壤TN的影响主要体现在大尺度上,且气温越高,土壤TN越低,这与Cao等^[33]和Geng等^[34]研究结果一致。在自然生态系统中,土壤中的矿质化过程和腐殖化过程共同影响着土壤TN的积累,在矿质化过程中,土壤微生物分解复杂有机物,而在腐殖化过程中,土壤微生物将合成较稳定的腐殖质并积累在土壤中。气温的升高,会增强土壤微生物的活性,加速有机质的矿化过程,导致土壤全氮含量减少^[35]。在区域尺度上,Deng等^[36]也强调了气温对土壤氮循环的关键作用。同时,本研究区主要作物类型为粮食和油料作物,相比于蔬菜、麻类和糖料作物,种植粮食与油料作物的区域因其较低的氮施用量,而具有更低的面源污染风险,这也可能是导致人为因素影响较低的原因之一^[37-38]。综合以上影响因素,本研究结果进一步验证了气温在土壤TN分布中的主导作用。

MGWR建模结果表明,pH的带宽相对较小(48 km),这表明其影响在空间上具有较高的异质性,这种空间异质性也反映了土壤的异质性特征。湟水河流域地形复杂,山地起伏,不同区域的土壤pH可能代表了因母质、成土过程和微地形差异而表现出的空间变化。这种复杂的地形条件导致土壤pH在小尺度上变化剧烈,进而影响土壤TN的分布。Vasu等^[39]研究也表明,土壤pH具有范围小,空间自相关性中等的特点,往往呈斑块状镶嵌分布。

这提示我们,在研究复杂地形条件下的土壤pH时,需要采用较高分辨率的数据,以更准确地捕捉其空间变化特征。

尽管距城市密集区距离的特征重要性相对较低,但该因素在MGWR建模结果中仍表现出显著性。结果表明,受人类活动扰动越小的地区,土壤TN含量更高。这个结果可能是由于城市化过程中的土地利用变化和农业活动对氮素循环的影响。人类活动(如较高强度的耕作)可能破坏土壤结构,减少土壤TN含量^[40]。中等程度的带宽说明其影响范围主要集中在城市周边,也说明随着地区的发展和城市化进程的推进,人类活动的影响正逐步显现。在本研究中,其他驱动因素具有较低的解释力,这是由于土壤是一个复杂的综合体,受到气候、植被、时间、生物、母质、地形和人为因素的综合作用^[41],而每个研究区具有不同的主导因素。同时,气候、人为因素等也可能解释或掩盖了其他因素的影响。

3.2 混合模型与原始模型、随机森林对比

本研究利用随机森林和皮尔逊相关性分析进行变量筛选,结合MGWR构建了混合模型。通过比较可以发现,原始模型虽然精度略高于混合模型,但该模型存在更强的过拟合效应,这与Shabrina等^[42]的研究结果一致。原始模型的 R^2 和有效 R^2 参数差距较大,这表明模型在训练数据上表现优异,但在测试数据上可能无法保持同样的性能。原始模型虽然具有较高的 R^2 ,但其AICc值劣于混合模型,这进一步证实了原始模型可能过度拟合了训练数据的噪声和细节,而非数据的真实结构。此外,本研究发现,在较多变量的情况下,MGWR模型均能拟合出一个相对较高的 R^2 ,然而这种高 R^2 很可能会牺牲模型的可解释性。相比之下,混合模型通过随机森林、皮尔逊相关性分析进行筛选,在减小了因素选取的主观性,降低了模型的复杂度,提高了模型的泛化能力,有利于进一步验证并分析各因素的尺度和空间上的影响。

随机森林虽然可以直接得出各因素的特征重要性,但是并不能分析因素在空间上的特征。此外,随机森林往往在训练集上表现出较高精度,但在交叉验证集和测试集上的精度低。表明随机森林能对训练数据进行较高精度的拟合,但在处理未见数据时表现较差。Verikas等^[20]也指出,随机森林

虽然精度较高,但对于变量的解释不足,重要性排序结果需要斟酌。而MGWR在处理空间数据方面表现出色,能够更准确地反映局部变量的影响,国内已有许多学者利用MGWR对土壤属性如土壤有机碳、土壤pH等展开研究^[25-26],这也是后续采用MGWR模型进行建模分析验证的原因。但是,MGWR本身无法处理变量之间的共线性,且在计算过高维的数据时容易出现误判,这将带来变量遗漏或变量选取较为主观的问题。本研究构建的混合模型可以更加客观、有效地进行变量筛选,处理具有空间异质性的数据。

4 结 论

为探究土壤氮素的驱动因素,维护湟水河流域生态脆弱区的生态系统稳定,本研究以湟水河流域为研究对象,探究了土壤TN的空间分布及驱动因素,并对土壤TN驱动因素模型进行了改进。主要结论如下:

1) 湟水河流域土壤TN呈现出较为显著的空间分布差异。研究区土壤TN含量呈现由东南向西北递增的趋势,且高值区主要分布在流域西北的高海拔地区,低值区零星分布在湟水河中下游段的两侧,总体呈现出带状分布的空间格局。

2) 湟水河流域土壤TN空间分布的驱动因素分为自然因素和人为因素。其中,年均气温是影响土壤TN的最主要驱动因素,且影响主要体现在较大尺度上;土壤pH与土壤TN呈负相关,但空间异质性较强;距城市密集区距离与土壤TN呈正相关,城市周边区域受人类活动影响较大。

3) 湟水河流域土壤TN的驱动因素预测模型改进。结合随机森林、皮尔逊相关性分析和MGWR模型的混合模型,能够减小因素选取的主观性和模型的复杂度,提高模型的泛化能力和可解释性,在处理空间数据方面表现优良。

同时,本研究也存在一定的局限性。例如采样点分布不够均匀,这可能对研究结果的精度产生一定影响。对于其他人类活动因素的定量描述较为困难,未来研究可以考虑通过大数据等手段来收集和整合多源数据,更全面地捕捉人类活动的影响。模型构建仍有提升空间,需要进一步优化和探索其他算法的应用,以提升模型的泛化能力和可解释性。

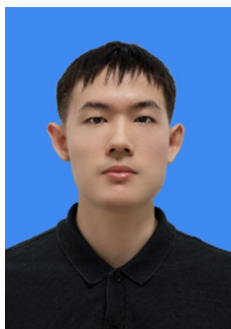
参考文献 References

- [1] Batjes N H. Total carbon and nitrogen in the soils of the world[J]. *European Journal of Soil Science*, 1996, 47(2): 151-163
- [2] Chen H, Ju P, Zhu Q, Xu X L, Wu N, Gao Y H, Feng X J, Tian J Q, Niu S L, Zhang Y J, Peng C H, Wang Y F. Carbon and nitrogen cycling on the Qinghai-Tibetan Plateau[J]. *Nature Reviews Earth & Environment*, 2022, 3(10): 701-716
- [3] Liu Z P, Shao M A, Wang Y Q. Spatial patterns of soil total nitrogen and soil total phosphorus across the entire Loess Plateau region of China[J]. *Geoderma*, 2013, 197: 67-78
- [4] Dempster A P, Schatzoff M, Wermuth N. A Simulation study of alternatives to ordinary least squares[J]. *Journal of the American Statistical Association*, 1977, 72(357): 77-91
- [5] 邹乐强. 最小二乘法原理及其简单应用[J]. 科技信息, 2010(23): 282-283
Zou L Q. Principle of least square method and its simple application[J]. *Science & Technology Information*, 2010(23): 282-283 (in Chinese)
- [6] 钟大放. 以加权最小二乘法建立生物分析标准曲线的若干问题[J]. 药物分析杂志, 1996(5): 343-346
Zhong D F. Some problems in establishing standard curve of bioanalysis by weighted least square method[J]. *Chinese Journal of Pharmaceutical Analysis*, 1996(5): 343-346 (in Chinese)
- [7] 田生昌. 最小二乘法的统计学原理及在农业试验分析中的应用[J]. 数学的实践与认识, 2015, 45(4): 124-133
Tian S C. The least squares method of statistical principles and applications in agricultural pilot study[J]. *Mathematics in Practice and Theory*, 2015, 45(4): 124-133 (in Chinese)
- [8] 郭立群. 数据区对最小二乘法拟合函数系数偏差的影响[J]. 物理实验, 1993(2): 70-72
Guo L Q. Influence of data area on coefficient deviation of least square fitting function[J]. *Physics Experimentation*, 1993(2): 70-72 (in Chinese)
- [9] 方匡南, 吴见彬, 朱建平, 谢邦昌. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38
Fang K N, Wu J B, Zhu J P, Xie B C. A review of technologies on random forests[J]. *Statistics & Information Forum*, 2011, 26(3): 32-38 (in Chinese)
- [10] 仇皓雷, 王海燕. 机器学习在土壤性质预测研究中的应用进展[J]. 生态学杂志, 2025, 44(1): 283-294
Qiu H L, Wang H Y. Application of machine learning to the prediction of soil properties: A review[J]. *Chinese Journal of Ecology*, 2025, 44(1): 283-294 (in Chinese)
- [11] Breiman L. Random forests[J]. *Machine Learning*, 2001, 45(1): 5-32
- [12] 卢宏亮, 赵明松, 刘斌寅, 张平, 陆龙妹. 基于随机森林模型的安徽省土壤属性空间分布预测[J]. 土壤, 2019, 51(3): 602-608
Lu H L, Zhao M S, Liu B Y, Zhang P, Lu L M. Spatial prediction of soil properties based on random forest model in Anhui Province[J]. *Soils*, 2019, 51(3): 602-608 (in Chinese)
- [13] 高文静, 夏冰, 芦园园, 应蓉蓉, 胡鹏杰, 李轶滑, 陈红枫, 钱家忠. 基于土壤剖面发生层和环境变量的安徽省土壤有机碳空间分布特征[J]. 土壤学报, 2025, 62(3): 677-691
Gao W J, Xia B, Lu Y Y, Hu P J, Li Y X, Chen H F, Qian J Z. Characteristics of spatial distribution of soil organic carbon in Anhui Province based on soil profile occurrence layer and environmental variables[J]. *Acta Pedologica Sinica*, 2025, 62(3): 677-691 (in Chinese)
- [14] 帅爽, 张志, 张天, 张焜, 肖成志, 陈安. 特征优化结合随机森林算法的干旱区植被高光谱遥感分类方法[J]. 农业工程学报, 2023, 39(9): 287-293

- Shuai S, Zhang Z, Zhang T, Zhang K, Xiao C Z, Chen A. Hyperspectral image classification method for dryland vegetation by combining feature optimization and random forest algorithm [J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2023, 39(9): 287-293 (in Chinese)
- [15] 包青岭, 丁建丽, 王敬哲. 利用随机森林方法优选光谱特征预测土壤水分含量[J]. 激光与光电子学进展, 2018, 55(11): 113002
- Bao Q L, Ding J L, Wang J Z. Prediction of soil moisture content by selecting spectral characteristics using random forest method [J]. *Laser & Optoelectronics Progress*, 2018, 55(11): 113 002 (in Chinese)
- [16] 王晗, 向友珍, 李汪洋, 史鸿棹, 王辛, 赵笑. 基于无人机多光谱遥感的冬油菜地上部生物量估算[J]. 农业机械学报, 2023, 54(8): 218-229
- Wang H, Xiang Y Z, Li W Y, Shi H Z, Wang X, Zhao X. Estimation of winter rapeseed above-ground biomass based on UAV multi-spectral remote sensing[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2023, 54(8): 218-229 (in Chinese)
- [17] 刘舒, 姜琦刚, 马玥, 肖艳, 李远华, 崔璨. 基于多目标遗传随机森林特征选择的面向对象湿地分类[J]. 农业机械学报, 2017, 48(1): 119-127
- Liu S, Jiang Q G, Ma Y, Li Y H, Cui C. Object-oriented wetland classification based on hybrid feature selection method combining with relief F, multi-objective genetic algorithm and random forest[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2017, 48(1): 119-127 (in Chinese)
- [18] 陈元鹏, 张世文, 罗明, 郎文聚, 鞠正山, 李少帅. 基于高光谱反演的复垦区土壤重金属含量经验模型优选[J]. 农业机械学报, 2019, 50(1): 170-179
- Chen Y P, Zhang S W, Luo M, Yun W J, Ju Z S, Li S S. Empirical model optimization of hyperspectral inversion of heavy metal content in reclamation area[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2019, 50(1): 170-179 (in Chinese)
- [19] 宋欠欠, 李轶群, 侯艳, 李康. 随机森林的变量捕获方法在高维数据变量筛选中的应用[J]. 中国卫生统计, 2015, 32(1): 49-53
- Song Q Q, Li Y Q, Hou Y, Li K. The application of a random forest-based variable hunting method to variable selection in high-dimensional data[J]. *Chinese Journal of Health Statistics*, 2015, 32(1): 49-53 (in Chinese)
- [20] Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: A survey and results of new tests[J]. *Pattern Recognition*, 2011, 44(2): 330-349
- [21] Fotheringham A S, Brunson C, Charlton M. Geographically weighted regression[J]. *The Sage Handbook of Spatial Analysis*, 2009, 1: 243-254
- [22] Fotheringham A S, Yang W, Kang W. Multiscale geographically weighted regression (MGWR) [J]. *Annals of the American Association of Geographers*, 2017, 107(6): 1247-1265
- [23] Yu H, Fotheringham A S, Li Z, Oshan T, Kang W, Wolf L J. Inference in multiscale geographically weighted regression[J]. *Geographical Analysis*, 2020, 52(1): 87-106
- [24] 沈体雁, 于瀚辰, 周麟, 古恒宇, 何泓浩. 北京市二手住宅价格影响机制: 基于多尺度地理加权回归模型(MGWR)的研究[J]. 经济地理, 2020, 40(3): 75-83
- Shen T Y, Yu H C, Zhou L, Gu H Y, He H H. On hedonic price of second-hand houses in Beijing based on multi-scale geographically weighted regression: Scale law of spatial heterogeneity[J]. *Economic Geography*, 2020, 40(3): 75-83 (in Chinese)
- [25] 南富森, 李宗省, 张小平, 崔乔, 李玉辰, 熊雪婷, 杨雪杰, 杨安乐. 基于MGWR的黄河流域土壤有机碳密度和影响因素分析[J]. 环境科学, 2023, 44(2): 912-923
- Nan F S, Li Z S, Zhang X P, Cui Q, Li Y C, Xiong X T, Yang X J, Yang A L. Spatial distribution characteristics and influencing factors of soil organic carbon density in Yellow River Basin based on MGWR model [J]. *Environmental Science*, 2023, 44(2): 912-923 (in Chinese)
- [26] 赵明松, 陈宣强, 徐少杰, 邱士其, 王世航. 基于MGWR的土壤pH值空间建模及其影响因素分析[J]. 环境科学, 2023, 44(12): 6909-6920
- Zhao M S, Chen X Q, Xu S J, Qiu S Q, Wang S H. Spatial prediction modeling for soil pH based on multiscale geographical weighted regression (MGWR) and its influencing factors[J]. *Environmental Science*, 2023, 44(12): 6909-6920 (in Chinese)
- [27] 张调风, 朱西德, 王永剑, 李红梅, 刘彩红. 气候变化和人类活动对湟水河流域径流量影响的定量评估[J]. 资源科学, 2014, 36(11): 2256-2262
- Zhang T F, Zhu X D, Wang Y J, Li H M, Liu C H. The impact of climate variability and human activity on runoff changes in the Huangshui River Basin [J]. *Resources Science*, 2014, 36(11): 2256-2262 (in Chinese)
- [28] 周莉, 李保国, 周广胜. 土壤有机碳的主导影响因素及其研究进展[J]. 地球科学进展, 2005(1): 99-105
- Zhou L, Li B G, Zhou G S. Advances in controlling factors of soil organic carbon[J]. *Advances in Earth Science*, 2005(1): 99-105 (in Chinese)
- [29] 张艳娇. 湟水河流域农业面源污染与经济发展关系的实证研究[D]. 西宁: 青海大学, 2019
- Zhang Y J. Empirical study on the relationship between agricultural non-point source pollution and economic development in the Huangshui River Basin[D]. Xining: Qinghai University, 2019 (in Chinese)
- [30] 代子俊. 近30年湟水流域土壤养分时空变异及影响因素[D]. 西宁: 青海师范大学, 2018
- Dai Z J. Temporal and spatial variation of soil nutrients and influencing factors in the Huangshui River Basin over the past 30 years[D]. Xining: Qinghai Normal University, 2018 (in Chinese)
- [31] 鲍士旦. 土壤农化分析[M]. 第3版. 北京: 中国农业出版社, 2000
- Bao S D. *Soil and Agricultural Chemistry Analysis*[M]. 3rd ed. Beijing: China Agriculture Press, 2000 (in Chinese)
- [32] Krivoruchko K. Empirical bayesian kriging[J]. *ArcUser Fall*, 2012, 6(10): 1 145
- [33] Cao Y Z, Wang X D, Lu X Y, Yan Y, Fan J H. Soil organic carbon and nutrients along an alpine grassland transect across Northern Tibet[J]. *Journal of Mountain Science*, 2013, 10(4): 564-573
- [34] Geng Y, Baumann F, Song C, Zhang M, Shi Y, Kühn P, Scholten T, He J S. Increasing temperature reduces the coupling between available nitrogen and phosphorus in soils of Chinese grasslands[J/OL]. *Scientific Reports*, 2017, 7(1): 43524
- [35] Nie X Q, Xiong F, Yang L, Li C B, Zhou G Y. Soil nitrogen storage, distribution, and associated controlling factors in the northeast Tibetan Plateau shrublands[J]. *Forests*, 2017, 8(11): 416
- [36] Deng L, Peng C H, Kim D G, Li J W, Liu Y L, Hai X Y, Liu Q Y, Huang C B, Shangguan Z P, Kuzyakov Y. Drought effects on soil carbon and nitrogen dynamics in global natural ecosystems [J/OL]. *Earth-Science Reviews*, 2021214: 103501
- [37] 杨世琦, 颜鑫. 基于农作物施肥量视角的区域尺度农田面源污染风险评价研究[J]. 中国农业大学学报, 2023, 28(2): 147-159
- Yang S Q, Yan X. Risk assessment of non-point sources pollution in farmland based on main crop fertilizer application rate[J]. *Journal of China Agricultural University*, 2023, 28(2): 147-159 (in Chinese)
- [38] Guan Y, Zhou W, Bai Z, Cao Y, Huang Y, Huang H. Soil nutrient variations among different land use types after reclamation in the Pingshuo opencast coal mine on the Loess Plateau, China[J]. *Catena*, 2020, 188: 104427
- [39] Vasu D, Singh S K, Sahu N, Tiwary P, Chandran P, Duraisami V P, Ramamurthy V, Lalitha M, Kalaiselvi B. Assessment of spatial variability

- of soil properties using geospatial techniques for farm level nutrient management[J]. *Soil and Tillage Research*, 2017, 169: 25-34
- [40] Wang J, Wang K L, Zhang M Y, Zhang C H. Impacts of climate change and human activities on vegetation cover in hilly Southern China[J]. *Ecological Engineering*, 2015, 81: 451-461
- [41] Bockheim J G, Gennadiyev A N, Hartemink A E, Brevik E C. Soil-forming factors and soil taxonomy[J]. *Geoderma*, 2014, 226: 231-237
- [42] Shabrina Z, Buyuklieva B, Ng M K M. Short-term rental platform in the urban tourism context: A geographically weighted regression (GWR) and a multiscale GWR (MGWR) approaches[J]. *Geographical Analysis*, 2021, 53(4): 686-707

责任编辑: 杨爱东



第一作者简介: 万成瑞, 博士研究生, 中国农业大学土地科学与技术学院土壤学专业。本科毕业于江西农业大学, 曾获得学业一等奖学金; 在中国农业大学硕博连读期间获得学业二等奖学金。先后参与国家自然科学基金项目 3 项, 参与专利 2 项。参与自然资源部农用地质量与监控重点实验室学术报告会并获银奖。作为第一作者在 *Environmental Research* 上发表论文 1 篇, 参与编写学术专著 1 部。



通讯作者简介: 黄元仿, 教授, 博士生导师。国际土壤学联合会土壤工程与技术委员会主席, 中国土壤学会副理事长, 北京土壤学会理事长。主要从事土壤过程量化、耕地保育、土地整治与生态修复等研究。主持国家自然科学基金、国家重点研发计划等项目课题 30 余项; 在《农业工程学报》、*Environmental Science and Technology*、*Geoderma* 等期刊发表学术论文 200 余篇; 出版学术专著 30 余部。